



(11)

EP 2 068 511 A1

(12)

EUROPEAN PATENT APPLICATION

(43) Date of publication:
10.06.2009 Bulletin 2009/24

(51) Int Cl.:
H04L 12/56 (2006.01)

(21) Application number: **07254729.2**

(22) Date of filing: **06.12.2007**

(84) Designated Contracting States:
**AT BE BG CH CY CZ DE DK EE ES FI FR GB GR
HU IE IS IT LI LT LU LV MC MT NL PL PT RO SE
SI SK TR**
Designated Extension States:
AL BA HR MK RS

(71) Applicant: **LUCENT TECHNOLOGIES INC.**
Murray Hill NJ 07974-0636 (US)

(72) Inventor: **Richa, Malhotra**
Overijssel (NL)

(74) Representative: **Cockayne, Gillian**
Alcatel-Lucent Telecom Limited
Unit 18, Core 3,
Workzone
Innova Business Park
Electric Avenue
Enfield
EN3 7XU (GB)

(54) Controlling congestion in a packet switched data network

(57) A method of controlling congestion in a packet switched network includes transmitting a pause message to a node upstream of a congested node to instruct it to cease sending for a period of time. The congested node also sends information to the upstream node informing it of the buffer size required to relieve congestion. If the buffer capacity of the upstream node is greater than or

equal to the signalled amount required, no further action is taken (7). If the buffer capacity of the upstream node is less than the required amount (6), it in turn sends a pause message to a node upstream of it, together with an indication of the required buffer size to reduce congestion (8), taking into account the amount available from the sending nodes.

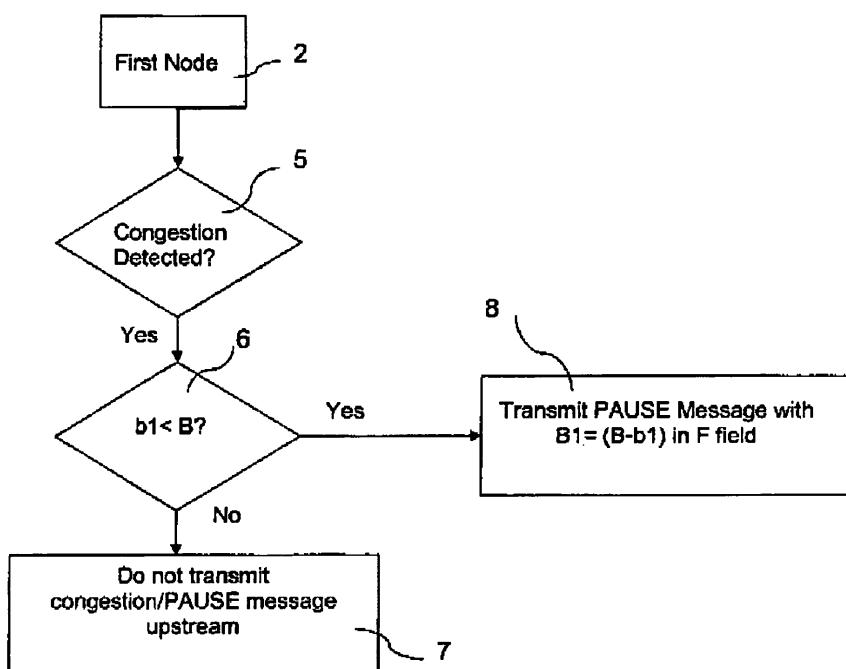


Fig. 2

Description

FIELD OF THE INVENTION

[0001] The present invention relates to a method of controlling congestion in a packet switched data network and to a network using such a method.

BACKGROUND OF THE INVENTION

[0002] Optimization of buffer sizes in routers, switches or bridges is difficult to achieve. Excessively large buffers lead to higher costs for those network elements that include them and also to performance issue such as high latency. In contrast, small buffers are cheaper but also increase packet loss for unpredictable bursty traffic. For higher layer applications like Transmission Control Protocol ("TCP"), used, for example, for file transfers and web-browsing, packet loss triggers a reduction in traffic sending rate. This in turn causes lower throughput and greater file download and response times. The effect of buffer size on TCP performance is dependent on a number of parameters and on the specific circumstances, taking into account traffic and network configuration. A paper by D. Wischik and N. McKeown "Part I: Buffer Sizes for Core Routers", ACM SIGCOMM Computer Communication Review, Volume 35, Number 2, July 2005. hereby incorporated by reference, discusses some of the considerations involved in determining suitable buffer sizes.

[0003] Ethernet back-pressure is a hop-by-hop congestion control mechanism. In time of congestion, where a node has insufficient buffer capability to accommodate all the traffic that is being sent, the congested node can signal its upstream neighbour to stop transmitting data using a PAUSE message, as set out in IEEE Standard 802.3-2005 at 31B.3 "Detailed specification of PAUSE operation". The PAUSE message instructs the sending upstream node to stop transmitting for set period of time. If, as a result of temporarily stopping its transmission, the upstream node also becomes congested, it can also transmit a PAUSE message to the node upstream from it, and so on. Thus, when there is congestion, it may lead to utilizing buffers at multiple nodes to hold packets. There is a risk that, under some conditions, utilizing the PAUSE mechanism may jam an entire network as the PAUSE instruction propagates upstream over multiple nodes.

BRIEF SUMMARY OF THE INVENTION

[0004] According to a first aspect of the invention, a method of controlling congestion in a packet switched data network, comprises the steps of: detecting congestion at a first node of the network; when congestion is detected, sending a pause message to a second node, upstream from the first node, to stop transmitting data to the first node for a period of time, and sending information to the second node indicating the amount of buffer space

required to relieve congestion; and, where the buffer capacity of the second node is less than that required, acting to reduce the congestion. The pause message may be a PAUSE message in accordance with Ethernet standards, but the invention is applicable to networks using other operating protocols. It is only necessary for the widest scope of the invention that the pause message results in the transmission of data from a sending node to be halted for a period of time. Use of the invention provides

5 a way to control the Ethernet back-pressure congestion control mechanism and also to provide flexible buffer space for nodes in a packet-switched network.

[0005] The second node may drop packets to reduce congestion. In another variant in accordance with the invention, where the buffer capacity of the second node is less than that required, a pause message is sent to a third node, upstream from the second node, thus acting to reduce the congestion. Where the buffer capacity of the second node is also less than that required, information

10 may be sent to the third node indicating the amount of buffer space required taking into account the buffer capacity of the second node. As the required buffer size is modified depending on the capacity of a number of nodes involved, and a pause message is transmitted upstream only when a particular node in a chain is unable

15 to meet that required capacity, the pause message is less likely to spread in an uncontrolled manner. Thus, by using the invention, there is a reduced likelihood that the entire network will become jammed.

[0006] The pause messages may themselves also include information indicating the amount of buffer space required. Where the pause message is a PAUSE message in accordance with Ethernet protocols, it may include bits additional to the frame structure specified in IEEE 802.3x standard which are used to indicate the existence of congestion and the amount of buffer capacity required.

[0007] In another method in accordance with the invention, information indicating the amount of buffer space required is sent in a different type of message than a pause message. Thus, the pause message could be sent separately to the message indicating a required buffer capacity.

[0008] The existence of a congestion state at a node 45 may be explicitly or implicitly signalled to an upstream node. By sending a message with a buffer size requirement, either as part of the pause message or accompanying the pause message but separate therefrom, the upstream node may deduce that congestion has been noted at the downstream node. However, in some methods in accordance with the invention, the existence of congestion is explicitly set out, for example, by including message bits dedicated to congestion information. Such congestion information maybe merely an indication of 50 whether or not congestion exists. In other cases, the type of congestion involved may be included by appropriate selection of one from a number of congestion codes. For example, it maybe useful to indicate that many TCP con-

nections are causing the congestion and that they are synchronized, with packet loss affecting them all at the same time. Another type of congestion may occur where there is only partial synchronization. In another type, the presence of persistent flows may lead to congestion.

[0009] In another method in accordance with the invention, a management system may also be involved. The management system may have an overview of the network architecture, traffic flow and processing capability to assess the buffer size required to ease congestion at a congested node or nodes. A congested node may send a message to the management system indicating the existence of congestion, possibly also including an indicator of the type of congestion it is experiencing. The management system may then reply to the congested node with at least one of an indication of the buffer size and the network path required to relieve congestion, which may be dependent on the type of congestion concerned. The network path may, for example, be designated by the node IDs over which the congestion causing flow passes. In this case the pause message might be directed only along this path. The management system may also send an indication of to which upstream node the pause message and indication of buffer size should be sent by the congested node.

[0010] According to another aspect of the invention, a data network has a congestion control mechanism in accordance with the method disclosed above.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Some embodiments of the present invention will now be described by way of example only, and with reference to the accompanying drawings, in which:

Figure 1 schematically illustrates a data network in accordance with the invention.;
 Figure 2 schematically illustrates steps of a method in accordance with the invention associated with a first node of the network of Figure 1;
 Figure 3 schematically illustrates steps of a method in accordance with the invention associated with a second node of the network of Figure 1;
 Figure 4 schematically illustrates steps of another method in accordance with the invention associated with a first node of the network of Figure 1; and
 Figure 5 schematically illustrates another data network in accordance with the invention.

DETAILED DESCRIPTION OF THE DRAWINGS

[0012] With reference to Figure 1, a data network 1 operating in accordance with the Ethernet protocol includes a plurality of nodes 2, 3 and 4, each of which includes a buffer at which frames are temporarily stored. Each node has an indicator for detecting the amount of the buffer capacity that is occupied and the amount which is free and available to be used to hold incoming data

packets. The nodes maybe routers, switches, bridges, or other network elements where data packets are required to be held. The first node 2 is downstream of the second node 3, receiving data packets therefrom. Similarly, the second node 3 is downstream of a third node 4. In practice, a node may connect to more than one other node in an upstream or downstream direction.

[0013] The first node 2 receives frames, or data packets, from the second node 3. If these are transmitted at too great a rate, the first node 2 becomes congested because there is insufficient buffer space remaining at the first node 2. The existence of congestion maybe determined by a number of different methods. For example, it may be selected to exist when the first node is unable to receive another data packet because the buffer has insufficient space to accommodate even one additional packet, or the existence of congestion could be based on the rate of receipt of incoming packets, or by calculations based on traffic management principles, or if the queue occupancy exceeds a certain threshold or by measuring average queue occupancy and if this average exceeds a certain pre-specified threshold

[0014] When congestion is detected at the first node 2, as shown at 5 in Figure 2, the amount of buffer space B required to resolve the congestion condition is compared at 6 to the buffer capacity b1 of the first node 2. The value of B is selected so as to be optimized to deal with network congestion and is a fixed value. In other alternative methods in accordance with the invention the value of B is variable. If the buffer capacity b1 of the first node is greater than or equal to B, then no action is taken, as shown at 7. However, if the buffer capacity b1 is less than B, then a PAUSE message is transmitted to the second node 3 upstream of the first node 2, as shown at 8. The PAUSE message instructs the second node 3 to stop transmitting to the first node 2 for a fixed period of time. The PAUSE message is in accordance with IEEE 802.3-2005 and also includes additional bits in the F field to indicate a new value B1, where B1=B-b1.

[0015] The second node 3 receives the PAUSE message and reads it, as shown at 9 of Figure 3. It re-sets the value B to be B1, thus taking into account the capacity available at the first node 2. At 10. the second node compares the new value of B with its own buffer capacity b2. If the buffer capacity b2 is greater than, or equal to, B, then no further action is taken, as shown at 11. If, however, b2 is smaller than B, the second node then determines if it is congested at 12. If it is congested, it sends a PAUSE message upstream to the third node 4, including a new value for B1. that is B1= B-b2, in its F field, as shown at 13.

[0016] The mechanism is repeated for other upstream nodes until congestion is relieved.

[0017] In the method described above with reference to Figures 2 and 3, there is no specific mention that a congestion state exists, as this is indicated by the inclusion of a value for B1 in the PAUSE message. However, in other methods, an indication that there is congestion

is also transmitted with the PAUSE message. This can be useful where different types of congestion conditions may exist and it is desired to differentiate between them. The value of B1 may be selected depending on the type of congestion involved.

[0018] In another method in accordance with the invention, as shown in Figure 4, the PAUSE message transmitted at 14 contains no information relating to buffer size or congestion. A separate message is transmitted from the first node 2 to the second node 3 at 15, and this include the buffer size B and also in this particular method, a congestion type indicator C.

[0019] With reference to Figure 5, a data network 16 includes nodes 17, 18 and 19 and also a management system 20. In one method in accordance with the invention, the first node 17 sends a congestion alert message to the management system 20. The management system 20 has an overview of the network topology, traffic flow on its path and so on. The management system 20 signals back to the congested node the value of B, the buffer space required for optimal performance. The value of B may be adjusted to take into account the congestion characteristics at the congested node. The management system may also specify an upstream node to which a PAUSE message is to be sent by the first node 17. The management system 20 may also signal back to the first node 17 as to what type of congestion exists, for onward transmission.

[0020] Although in the example described above, reference is made to PAUSE messages in accordance with Ethernet protocols, it is not essential to the invention that the method is implemented in an Ethernet arrangement. The pause message may be of any suitable structure conveying information to a node concerning stopping transmission of data packets for a period of time. Alternatively, the management system might also specify the entire network path consisting of a string of nodes and their IDs over which the problem causing traffic flow passes. In this case the PAUSE message will be directed to only nodes on this specified path. In order to realize the method like this the management system could send a special message to the congested node with the IDs (Ethernet MAC addresses or IP addresses) of the nodes involved.

[0021] The present invention may be embodied in other specific forms, and carried out by other methods, without departing from its spirit or essential characteristics. The described embodiments and methods are to be considered in all respects only as illustrative and not restrictive. The scope of the invention is, therefore, indicated by the appended claims rather than by the foregoing description. All changes that come within the meaning and range of equivalency of the claims are to be embraced within their scope.

Claims

1. A method of controlling congestion in a packet switched data network, comprising the steps of:
5
detecting congestion at a first node of the network;
when congestion is detected, sending a pause message to a second node, upstream from the first node, to stop transmitting data to the first node for a period of time, and sending information to the second node indicating the amount of buffer space required to relieve congestion;
and, where the buffer capacity of the second node is less than that required, acting to reduce the congestion.
2. The method as claimed in claim 1 and wherein the second node drops packets to reduce congestion.
10
3. The method as claimed in claim 1 and, where the buffer capacity of the second node is less than that required, acting to reduce the congestion by sending a pause message to a third node, upstream from the second node.
15
4. The method as claimed in claim 3 and wherein, where the buffer capacity of the second node is less than that required, sending information to the third node indicating the amount of buffer space required taking into account the buffer capacity of the second node.
20
5. The method as claimed in any preceding claim and wherein the pause messages are PAUSE messages in accordance with Ethernet protocols.
25
6. The method as claimed in any preceding claim and wherein the pause messages include information indicating the amount of buffer space required.
30
7. The methods as claimed in claim 6 when dependent on claim 5 and wherein a PAUSE message frame in accordance with IEEES 802.3x includes additional bits used to indicate the existence of congestion and the amount of buffer capacity required.
35
8. The method as claimed in any preceding claim and wherein information indicating the amount of buffer space required is sent in a different type of message than a pause message.
40
9. The method as claimed in any preceding claim and wherein information indicating the existence of congestion is sent from the first node to a traffic management system and the traffic management system informs the first node of at least one of: the amount of buffer space required to relieve congestion and
55

the network path required to relieve congestion.

10. The method as claimed in claim 9 wherein, where the first node is informed of said network path, the pause messages are directed only along said network path. 5
11. The method as claimed in any preceding claim and wherein the amount of buffer space required to relieve congestion is dependent on the type of congestion. 10
12. A data network having a congestion control mechanism in accordance with the method of any of claims 1 to 11. 15

20

25

30

35

40

45

50

55

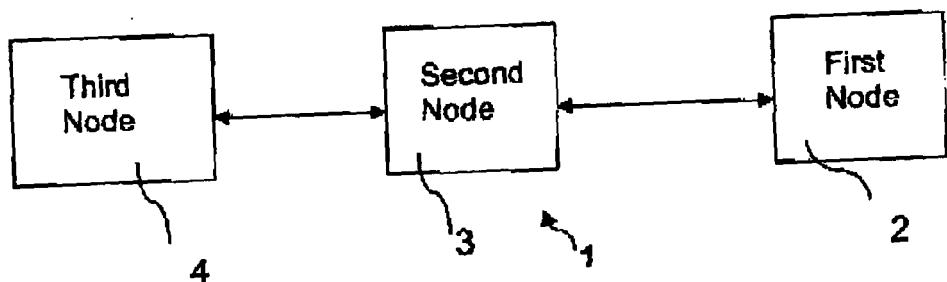


Fig. 1

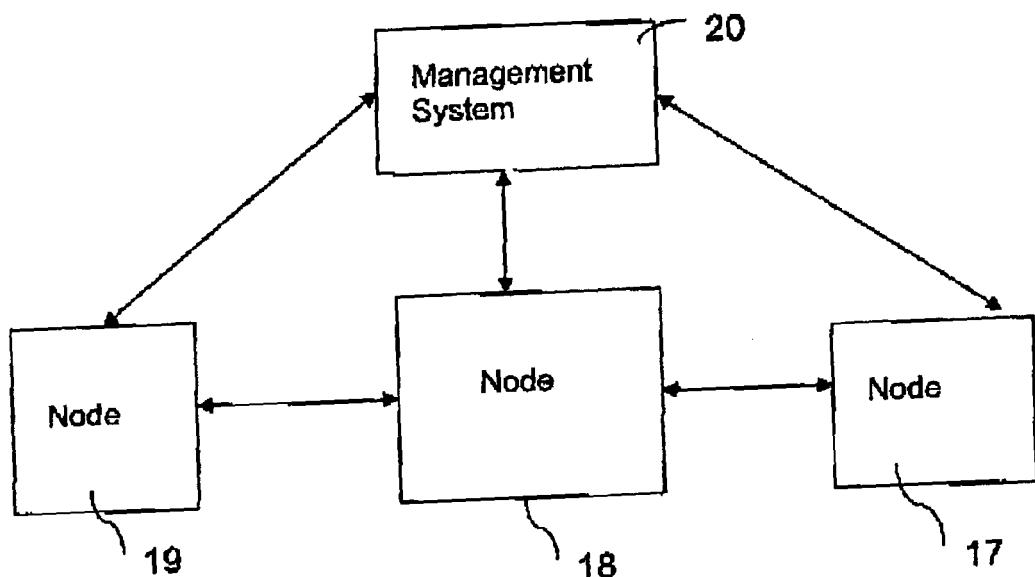
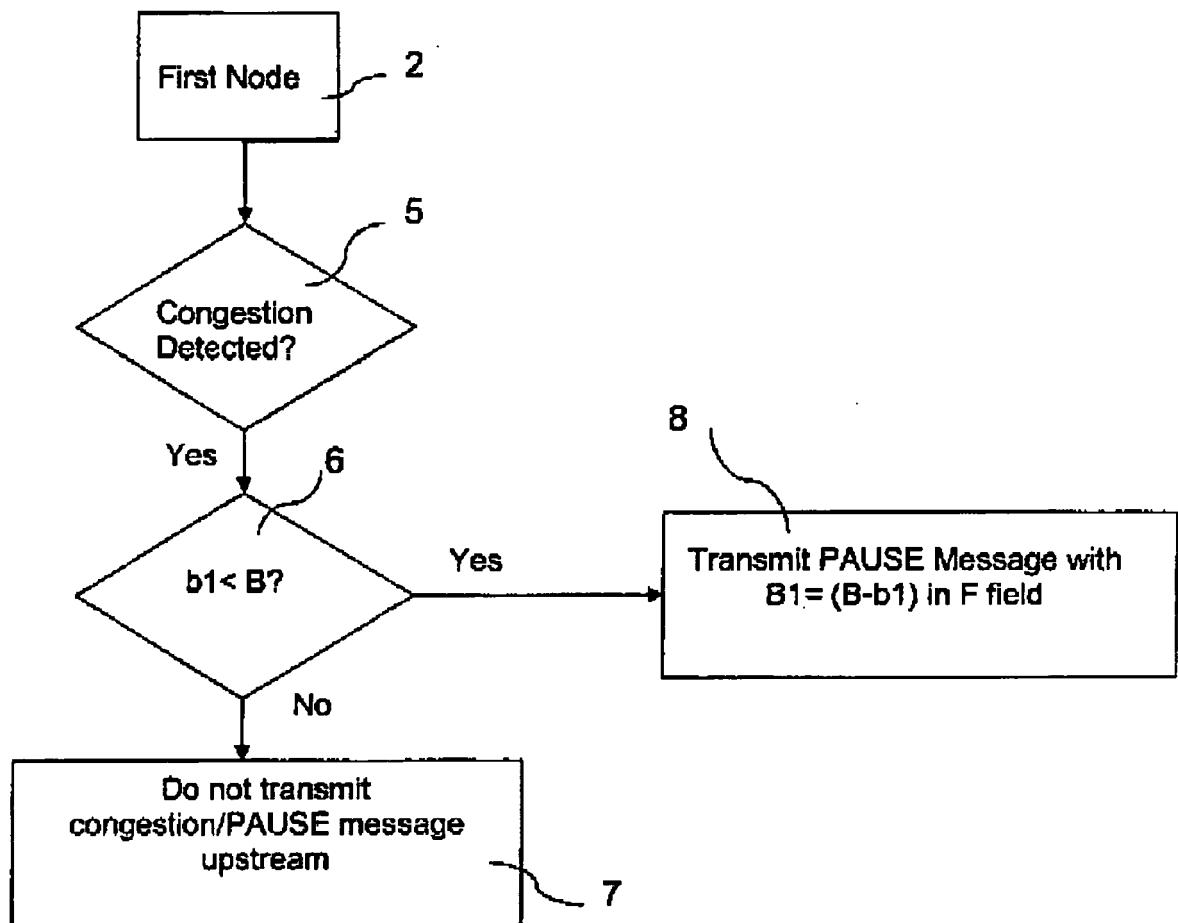


Fig. 5

**Fig. 2**

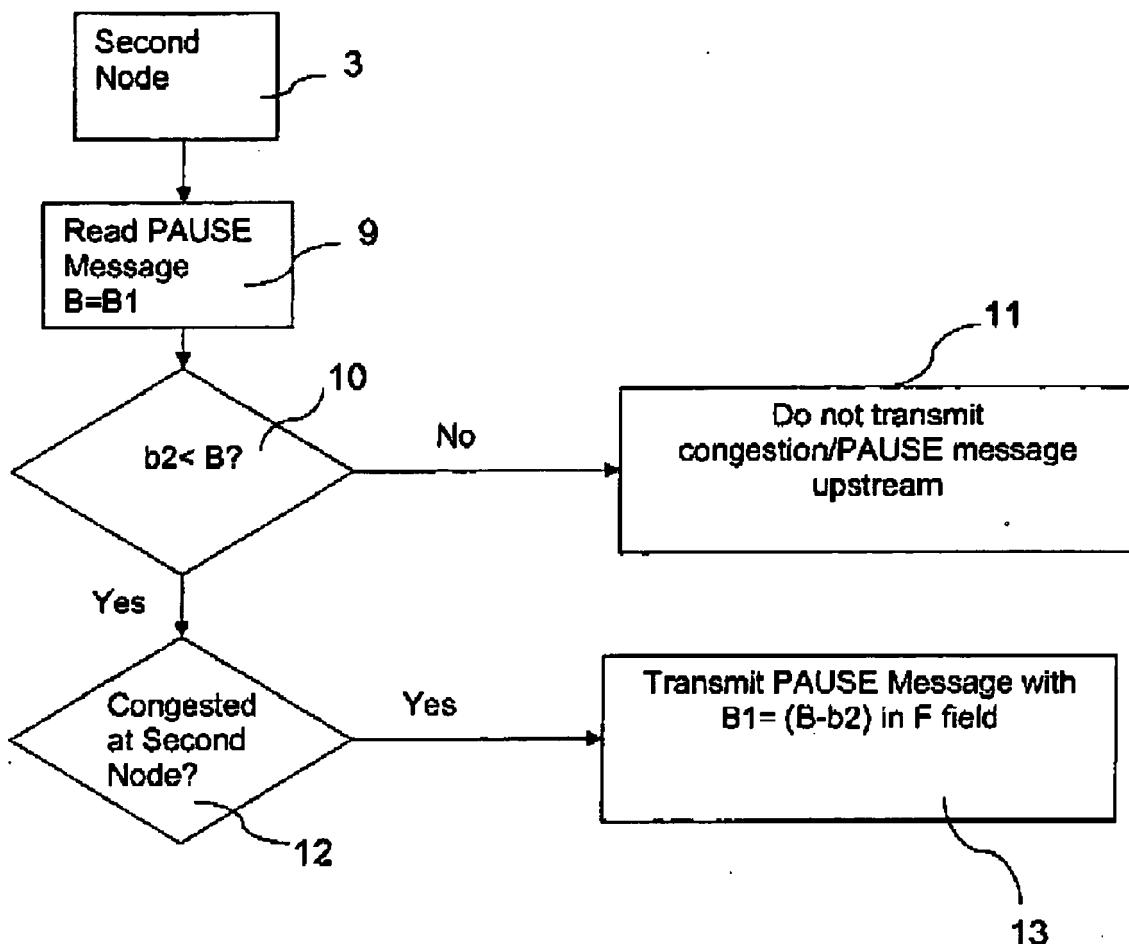


Fig. 3

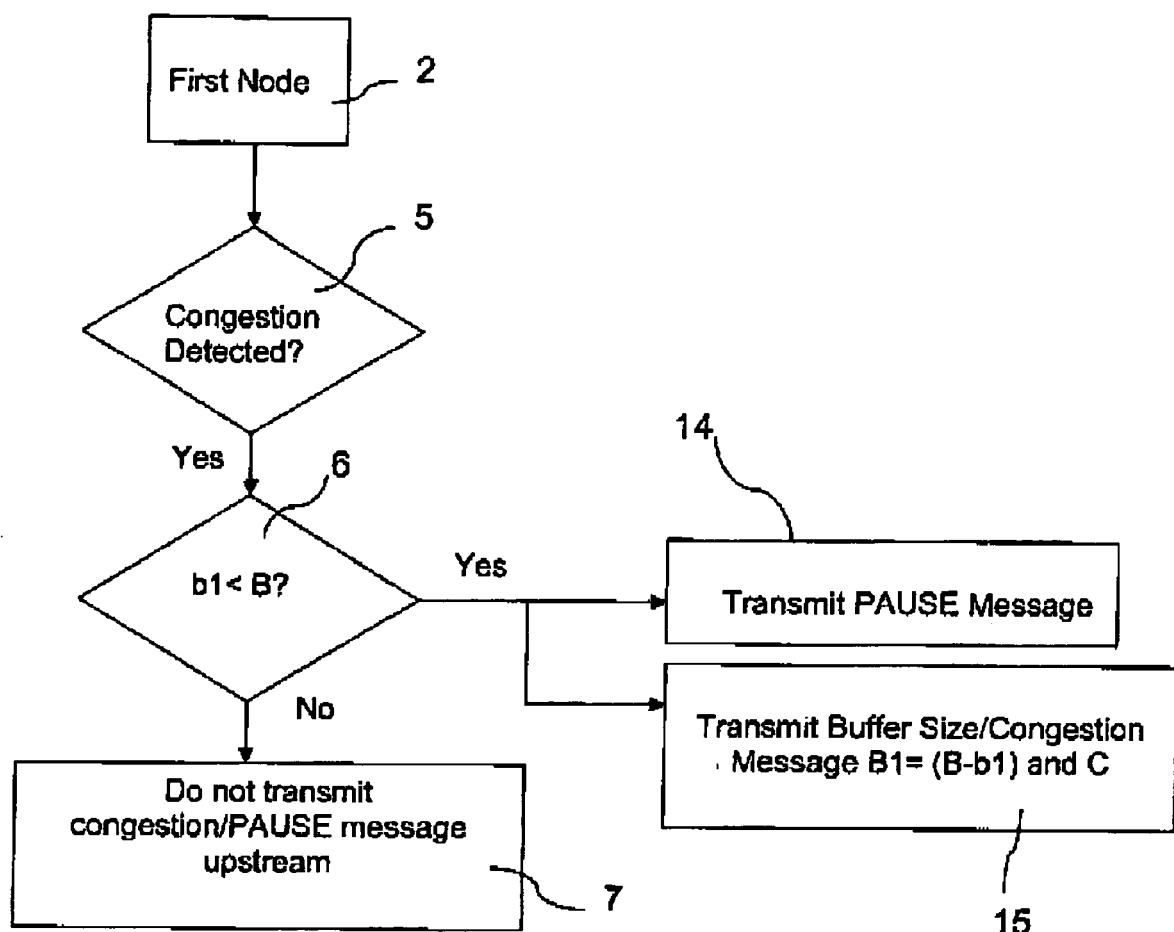


Fig. 4



DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	US 2002/136163 A1 (KAWAKAMI TETSUYA [JP] ET AL) 26 September 2002 (2002-09-26) * figures 5,8,9,17a,18,19,27 * * paragraph [0018] - paragraph [0028] * * paragraph [0108] - paragraph [0112] * * paragraph [0177] - paragraph [0219] * -----	1-5,12	INV. H04L12/56
Y	US 2005/243722 A1 (LIU ZHEN [US] ET AL) 3 November 2005 (2005-11-03) * paragraph [0036] - paragraph [0044] * * figures 1-3 *	6-11	
A	----- WO 2007/047865 A (AMPLE COMMUNICATIONS INC [US]; ELLEBRECHT EDWARD; THALKA MAREK [US]; P) 26 April 2007 (2007-04-26) * page 6, line 27 - page 7, line 9; figures 6a,7a *	1-5,12	
A	-----	1-12	
TECHNICAL FIELDS SEARCHED (IPC)			
H04L			
<p style="text-align: center;">The present search report has been drawn up for all claims</p> <p style="text-align: center;">2</p>			
<p>Place of search</p> <p>Munich</p>		<p>Date of completion of the search</p> <p>10 March 2008</p>	Examiner
<p>CATEGORY OF CITED DOCUMENTS</p> <p>X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document</p> <p>T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document</p>			

**ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.**

EP 07 25 4729

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on. The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

10-03-2008

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 2002136163	A1	26-09-2002	NONE		
US 2005243722	A1	03-11-2005	CA 2564363 A1	17-11-2005	
			CN 1965532 A	16-05-2007	
			EP 1747644 A2	31-01-2007	
			JP 2007535879 T	06-12-2007	
			KR 20070003983 A	05-01-2007	
			WO 2005109772 A2	17-11-2005	
WO 2007047865	A	26-04-2007	NONE		

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- **D.WISCHIK ; N. MCKEOWN.** Part I: Buffer Sizes for Core Routers. *ACM SIGCOMM Computer Communication Review*, July 2005, vol. 35 (2 [0002])